RA

stichting
mathematisch
centrum

MC

REKENAFDELING RA NR 19/71 JULI

W. HOFFMANN EN J.P. HOLLENBERG
FOUTENBEREKENING VOOR LINEAIRE KLEINSTE-KWADRATENPROBLEMEN

2e boerhaavestraat 49 amsterdam

Inhoud

Inleiding

Laten gegeven zijn een reële m × n matrix A, $m \geq n$ en een reële m-vector b.
Voor een willekeurige n-vector x definiëren we r = Ax-b. Onder de kleinste-kwadratenoplossing van [A:b] verstaan we die vector x waarvoor $||r||_2$ minimaal is; x voldoet aan $A^TAx = A^Tb$ (zie [1], p.57).

De 2-norm van een vector v (notatie: $||v||_2$) wordt gedefinieerd door:

$$||v||_2 = \sqrt{v^Tv} = \sqrt{\sum_i (v_i)^2}.$$

De 2-norm van een matrix M (ook spectraal norm genoemd) wordt gedefinieerd door:

$$||M||_2 = \sqrt{\lambda_{max}(M^TM)}.$$

Van belang is ook de Euclidische norm (oftewel Frobenius-norm) van een matrix:

$$||M||_E = \sqrt{\sum_i \sum_j (M_{ij})^2};$$

deze is namelijk makkelijker te berekenen dan de 2-norm.
Voor een n × n matrix M weten we:

$$||M||_2 \leq ||M||_E \leq \sqrt{n}\,||M||_2 .$$

Zie voor een verdere beschouwing over normen bijv. ([1], p. 7 e.v.)

In het vervolg zal met ||M||, resp. ||v|| altijd bedoeld worden $||M||_2$, resp. $||v||_2$.
Matrices en vectoren die in de oorspronkelijke probleemstelling voorkomen zullen we aangeven met $\tilde{M}$, resp. $\tilde{v}$.

De representatie van deze matrices en vectoren in de computer zullen we aangeven met M, resp. v.
Matrices en vectoren die een resultaat van bewerkingen in de computer zijn, worden aangegeven met $\bar{M}$, resp. $\bar{v}$.

In §1 geven we aan in hoeverre x verschilt van $\tilde{x}$; het is duidelijk dat dit onafhankelijk is van de gevolgde methode om $\bar{x}$ te berekenen.
In §2 geven we een foutenanalyse voor de "klassieke" methode om $\bar{x}$ te bepalen.
In §3 doen we dit voor de QR-methode volgens Householder.
In §4 tenslotte geven we aan welke verfijning aangebracht kan worden indien $\bar{r} = fl(A\bar{x}-b)$ berekend wordt.

## §1. Relatie met het oorspronkelijke probleem

Laat x de kleinste-kwadratenoplossing zijn van $[A\vdots b]$ en $\tilde{x}$ die van $[\tilde{A}\vdots\tilde{b}]$. Zij $\tilde{A}^T\tilde{A} = \tilde{M}$ en zij verder $A = \tilde{A} + \delta A$, $b = \tilde{b} + \delta\tilde{b}$. $\delta A$ en $\delta b$ stellen voor, de som van de waarnemingsfouten en de representatiefouten. De representatiefouten ontstaan door afronding tot in de computer representeerbare getallen en zijn gewoonlijk veel kleiner dan de waarnemingsfouten.
We weten dat geldt:

$$A^TAx = A^Tb,$$

hieruit volgt:

$$\tilde{A}^T\tilde{A}x + \tilde{A}^T\delta Ax + \delta A^TAx = \tilde{A}^T\tilde{b} + \tilde{A}^T\delta b + \delta A^Tb.$$

Dit levert ons:

1.1 $$\tilde{M}x = \tilde{A}^T\tilde{b} - \delta A^Tr - \tilde{A}^T(\delta Ax-\delta b).$$

Eveneens geldt:

1.2 $$\tilde{\tilde{M}}x = \tilde{A}^T\tilde{b}.$$

Uit combinatie van 1.1 en 1.2 volgt:

1.3 $$x - \tilde{x} = -\tilde{M}^{-1}\delta A^T r - \tilde{M}^{-1}\tilde{A}^T(\delta Ax - \delta b).$$

We weten dat $\tilde{A}$ geschreven kan worden als het product van een $m \times n$-orthonormale matrix $\tilde{Q}$, dwz. $\tilde{Q}^T\tilde{Q} = I$, en een $n \times n$ bovendriehoeks-matrix $\tilde{R}$. Hieruit volgt dat $\tilde{M} = \tilde{R}^T\tilde{R}$.
Toepassing hiervan in 1.3 levert ons:

1.4 $$x - \tilde{x} = -\tilde{M}^{-1}\delta A^T r - \tilde{R}^{-1}\tilde{Q}^T(\delta Ax - \delta b).$$

Dus geldt voor de norm hiervan wegens de driehoeksongelijkheid en de producteigenschap van normen (merk op dat $||\tilde{Q}^T|| = 1$):

1.5 $$||x-\tilde{x}|| \leq ||\tilde{M}^{-1}||\ ||\delta A^T||\ ||r|| + ||\tilde{R}^{-1}||\ (||\delta A||\ ||x|| + ||\delta b||).$$

Voor $||\tilde{M}^{-1}||$ kunnen we de volgende afschatting geven:

$$||\tilde{M}^{-1}|| \leq ||M^{-1}|| + ||\tilde{M}^{-1}||\ ||I - \tilde{M}M^{-1}||.$$

Indien $||I - \tilde{M}M^{-1}|| < 1$ is geldt dus

$$||\tilde{M}^{-1}|| \leq ||M^{-1}|| \ / \ (1 - ||I - \tilde{M}M^{-1}||).$$

Hieruit volgt:

1.6 $$||\tilde{M}^{-1}|| \leq ||\tilde{M}^{-1}|| /$$

$$\{1 - (||A^T||\ ||\delta A|| + ||\delta A^T||\ ||A|| + ||\delta A^T||\ ||\delta A||)||M^{-1}||\},$$

mits de noemer van deze laatste expressie positief is.

Analoog hieraan kunnen we voor $||\tilde{R}^{-1}||$ vinden:

1.7 $$||\tilde{R}^{-1}|| \leq ||R^{-1}|| \ / \ \{1 - ||R - \tilde{R}||\ ||R^{-1}||\},$$

mits ook in deze laatste expressie de noemer positief is. Hieraan is zeker voldaan als de noemer in 1.6 positief is.

Tot slot willen we over de betekenis van 1.5 het volgende opmerken:
Het rechterlid van 1.5 bestaat uit 2 termen waarvan de eerste $||\tilde{M}^{-1}||$ als factor heeft en de tweede $||\tilde{R}^{-1}||$. Daar $\tilde{M} = \tilde{R}^T\tilde{R}$ geldt ruwweg $||\tilde{M}^{-1}|| = O(||\tilde{R}^{-1}||^2)$.
Van een goed gesteld kleinste-kwadratenprobleem mogen we echter verwachten dat het residu r een kleine norm heeft, zodat $||r||$ helpt om de eerste term klein te krijgen.
Kwalitatief kunnen we nu zeggen dat $||x-\tilde{x}||$ niet te groot wordt, als $||\delta A||$ en $||\delta b||$ klein genoeg zijn, $||r||$ klein is en de conditie van het probleem niet al te slecht is, opdat $||\tilde{R}^{-1}||$ en $||\tilde{M}^{-1}||\ ||r||$ niet te groot worden.

## §2. Voorwaartse foutenanalyse van de "klassieke" methode

Zij $c = A^T b$, dan wordt het beschouwde probleem teruggebracht tot het oplossen van het vierkante stelsel:

$$Mx = c.$$

Veronderstel dat we werken met een floating-point arithmetiek met grondtal $\beta$ en een woordlengte van t digits.
We beschouwen de fouten in $\bar{M}$ en $\bar{c}$:

2.1 $\bar{M} = fl(A^TA) = A^TA + E,$

2.2 $\bar{c} = fl(A^Tb) = A^Tb + e.$

Er geldt (vgl. [1], blz. 22):

2.3 $$||E|| \leq m\, \beta^{-t_1} ||A^T||\ ||A||,$$

2.4 $$||e|| \leq m\, \beta^{-t_1} ||A^T||\ ||b||.$$

Hierin is $\beta^{-t_1} = 1.06C\beta^{-t}$ waarin C afhangt van de arithmetiek in de computer; indien $fl(a\times b)$ optimaal is, geldt $C = \beta/2$.

Voor een zekere $\hat{x}$ geldt:

2.5 $$\bar{M}\hat{x} = \bar{c}.$$

Dit is het stelsel dat in de computer opgelost wordt; de numeriek gevonden oplossing is gelijk aan $\bar{x}$.

Er bestaat een matrix F waarvoor geldt:

2.6 $$(\bar{M}+F)\bar{\bar{x}} = \bar{c}.$$

Indien 2.5 opgelost wordt met de methode van Cholesky kunnen we $||F||$ afschatten met:

2.7 $$||F|| \leq n^3 \beta^{-t_1}||\bar{M}||.$$

(Vergelijk [1], (5.5.2); doordat $\bar{M}$ positief definiet is, weten we dat de in (5.5.2) voorkomende f(n) gelijk is aan 1).

Combinatie van 2.1, 2.2 en 2.6 levert ons:

$$(M+E+F)\bar{\bar{x}} = c + e.$$

Hieruit volgt:

$$\bar{\bar{x}} - x = M^{-1}\{e-(E+F)\bar{\bar{x}}\},$$

of voor de norm hiervan:

2.8 $$||\bar{\bar{x}}-x|| \leq \beta^{-t_1}||M^{-1}||\{m||A^T|| \; ||b||+(m||A^T|| \; ||A||+n^3||\bar{M}||)||\bar{\bar{x}}||\}.$$

Indien $fl(\bar{\bar{M}}^{-1})$ berekend wordt, kunnen we een afschatting voor $||M^{-1}||$ geven:

$$M^{-1} = fl(\bar{\bar{M}}^{-1}) + M^{-1}\{I-Mfl(\bar{\bar{M}}^{-1})\}$$

$$= fl(\bar{\bar{M}}^{-1}) + M^{-1}\{I-\bar{M}fl(\bar{\bar{M}}^{-1}) + \bar{M}fl(\bar{\bar{M}}^{-1}) - Mfl(\bar{\bar{M}}^{-1})\}.$$

We weten dat ([1], blz. 40):

$$||\bar{\bar{M}}^{-1} - fl(\bar{\bar{M}}^{-1})|| \leq n^3 \beta^{-t_1}||fl(\bar{\bar{M}}^{-1})||$$

en:

$$\bar{M} - M = E.$$

Hieruit volgt

2.9 $$||M^{-1}|| \leq \frac{||fl(\bar{\bar{M}}^{-1})||}{1 - \beta^{-t_1}(n^3||\bar{M}|| + m||A^T|| \; ||A||)||fl(\bar{\bar{M}}^{-1})||}.$$

Combinatie van 2.9 en 2.8 levert ons de gewenste afschatting van $||\bar{\bar{x}}-x||$.

Laten we nu de samenhang tussen 2.8 en 1.5 bekijken. Stel daartoe

$$m\beta^{-t}||A^T|| = ||\delta\bar{A}^T||.$$

Er geldt dan:

2.10 $$||\bar{x}-x|| \leq ||M^{-1}||\ ||\delta\bar{A}^T||(||A||\ ||\bar{x}|| + ||b||) + n^3\beta^{-t}||M^{-1}||\ ||\bar{M}||\ ||\bar{x}||.$$

We zien als verschillen:

1°) in plaats van $||\tilde{R}^{-1}||\ ||\delta A||\ ||x||$ in 1.5 treedt $||M^{-1}||\ ||\delta\bar{M}||\ ||\bar{x}||$ in 2.10 op;

2°) in plaats van $||r||$ in 1.5 treedt $||A||\ ||\bar{x}|| + ||b||$ in 2.10 op;

3°) in plaats van de term $||\tilde{R}^{-1}||\ ||\delta b||$ in 1.5 treedt geen expliciete corresponderende term in 2.10 op; een equivalente bijdrage is in $||M^{-1}||\ ||\delta\bar{A}^T||\ ||b||$ verwerkt.

Van deze verschilpunten is dat onder 1° het belangrijkste.
Door dit verschil kan de bovengrens 2.10 te hoog uitvallen, ook als de gemaakte rekenfouten kleiner zijn dan $\delta A$ en $\delta b$ uit §1.
Al door de bepaling van $A^TA$ kunnen ernstige fouten geïntroduceerd worden. Een remedie hiertegen is de berekening van $A^TA$ en de Cholesky-ontbinding in dubbele-lengtearithmetiek uit te voeren.

## §3. Voorwaartse foutenanalyse van de QR-methode

Bij gegeven A bestaan een matrix Q en een matrix R waarvoor A = QR (zie §1). Het kleinste-kwadratenprobleem kunnen we reduceren tot:

3.1 $$Rx = Q^Tb = c.$$

De numeriek gevonden matrix $\bar{R}$ verschilt echter van R. Wel bestaat een orthogonale matrix $\hat{Q}$ waarvoor:

3.2 $$\bar{R} = \hat{Q}^T(A+E),$$

waarin

3.3 $$||E|| \leq 12.5\, n^{3/2}\, \beta^{-t_1}||A||, \qquad \text{(zie [2], [3])}.$$

Eveneens geldt:

3.4 $$\bar{c} = \hat{Q}^T(b+e),$$

waarin

3.5 $$||e|| \leq 12.5\, n\, \beta^{-t_1}||b||, \qquad \text{(zie [2], [3])}.$$

Zij $\hat{x}$ de exacte kleinste-kwadratenoplossing van $[A+E\,\vdots\,b+e]$ en x die van $[A\,\vdots\,b]$ dan geldt volgens 1.4:

3.6 $$x - \hat{x} = -M^{-1}\, E^T\, \hat{r} - R^{-1}\, Q^T(E\hat{x}-e),$$

waarbij:

$$\hat{r} = (A+E)\hat{x} - (b+e).$$

Daar we hier een foutenafschatting geven zonder residuberekening zullen we $||\hat{r}||$ afschatten met $||b|| + ||e||$. Er geldt namelijk voor $r = Ax - b$:

$$\begin{aligned} ||r||^2 = r^Tr &= x^TA^TAx - x^TA^Tb - b^TAx + b^Tb \\ &= x^T(A^TAx-A^Tb) - x^TA^Tb + b^Tb \\ &= b^Tb - x^TA^Tb \\ &= ||b||^2 - x^TA^TAx \\ &= ||b||^2 - ||Ax||^2 \leq ||b||^2. \end{aligned}$$

Voor de norm van $x - \hat{x}$ geldt dan:

3.7 $$\begin{aligned} ||x-\hat{x}|| \leq\ & ||M^{-1}||\ ||E^T||(||b|| + ||e||) + \\ & + ||R^{-1}||(||E||\ ||\hat{x}|| + ||e||). \end{aligned}$$

De berekende oplossingsvector $\bar{x}$ wordt verkregen door een terugsubstitutie toe te passen op het stelsel 3.1.
Er geldt voor zekere matrix F:

3.8 $$(\bar{R}+F)\bar{\bar{x}} = \bar{\hat{Rx}},$$

waarin (volgens [1], (5.4.3) met gebruikmaking van $||F||_2 \leq \sqrt{n}||F||_1$ (vgl. [1], pag. 10) en $g = ||\bar{R}||$ wegens $\bar{R}^T\bar{R}$ pos. def.):

3.9 $$||F|| \leq n^{5/2}||\bar{R}||\ \beta^{-t_1}.$$

Nu geldt:

3.10 $$||\bar{\bar{x}}-\hat{x}|| \leq ||\bar{\bar{R}}^{-1}||\ ||F||\ ||\bar{\bar{x}}||$$

en

3.11 $$||\hat{x}|| \leq (1 + ||\bar{\bar{R}}^{-1}||\ ||F||)||\bar{\bar{x}}||.$$

Uiteindelijk krijgen we dus als bovengrens voor $||\bar{\bar{x}}-x||$:

3.12 $$\begin{aligned} ||\bar{\bar{x}}-x|| \leq\ & ||\bar{\bar{R}}^{-1}||\ ||F||\ ||\bar{\bar{x}}|| \\ & + ||M^{-1}||\ ||E^T||(||b||+||e||) \\ & + ||R^{-1}||\{||E||(1 + ||\bar{\bar{R}}^{-1}||\ ||F||)||\bar{\bar{x}}||+||e||\}. \end{aligned}$$

Ter completering van 3.12 moeten we nog de volgende afschattingen geven:

3.13 $$||R^{-1}|| \leq \frac{||\bar{\bar{R}}^{-1}||}{1 - ||\bar{R}-R||\ ||\bar{\bar{R}}^{-1}||},$$

3.14 $$||\bar{\bar{R}}^{-1}|| \leq \frac{||fl(\bar{\bar{R}}^{-1})||}{1 - n^3\ \beta^{-t_1}||fl(\bar{\bar{R}}^{-1})||}.$$

Voor de afschatting van $||M^{-1}||$ zie 2.9.
In de praktijk geldt, als $n^3\ \beta^{-t_1}||fl(\bar{\bar{R}}^{-1})|| << 1$, dan is ook $||\bar{R}-R||\ ||\bar{\bar{R}}^{-1}|| << 1$, dus, blijkens formules 3.13 en 3.14, zijn dan $||R^{-1}||$ en $||\bar{\bar{R}}^{-1}||$ practisch even groot als de door berekening verkregen grootheid $||fl(\bar{\bar{R}}^{-1})||$.

## §4. Foutenanalyse met residuberekening

Als we een vector $\bar{\bar{x}}$ berekend hebben als benadering voor de oplossingsvector van een kleinste-kwadratenprobleem, dan kunnen we de residuvector $r = A\bar{\bar{x}}-b$

berekenen. We vinden:

4.1 $$\bar{r} = fl(A\bar{x}-b) = A\bar{x} - b + \delta r,$$

met:

4.2 $$||\delta r|| \leq \{(n+1)\ ||A||\ ||\bar{x}|| + ||b||\}\beta^{-t_1}.$$

Zij verder $s = A^T\bar{r}$, dan geldt:

4.3 $$\bar{s} = fl(A^T\bar{r}) = A^T\bar{r} + \delta s,$$

met:

4.4 $$||\delta s|| \leq m\ \beta^{-t_1}||A^T||\ ||\bar{r}||.$$

Uit 4.1 volgt:

4.5 $$A^T\bar{r} = A^TA\bar{x} - A^Tb + A^T\delta r.$$

Uit 4.5, 4.3 en de relatie $A^TAx = A^Tb$, volgt:

4.6 $$\bar{x} - x = M^{-1}\{\bar{s}-\delta s-A^T\delta r\},$$

dus geldt voor de norm:

4.7 $$||\bar{x}-x|| \leq ||M^{-1}||\{||\bar{s}|| + m\beta^{-t_1}||A^T||\ ||\bar{r}||\}$$
$$+ ||R^{-1}||\beta^{-t_1}\{(n+1)\ ||A||\ ||\bar{x}|| + ||b||\}.$$

De afschatting voor $||M^{-1}||$ is gegeven in 2.9, die voor $||R^{-1}||$ in 3.13.

Indien $\bar{r}$ en $\bar{s}$ in dubbele-lengteprecisie worden berekend, geldt bij benadering:

4.8 $$||\bar{x}-x|| \leq ||M^{-1}||\ ||\bar{s}||.$$

Literatuur:

1. T.J. Dekker, Numerieke Algebra.
MC Syllabus 12, Mathematisch Centrum, Amsterdam 1971.

2. Å. Björck, Iterative refinement of linear least squares solutions I,
BIT 7, (1967) 257-278.

3. G.H. Golub & J.H. Wilkinson,
Note on the iterative refinement of least squares solution,
Num. Mat. 9 (1966) 139-148.